TR-IT-0260

Repair Strategies for German Generation in TDMT — Problems and Prospects —

Julia E. Heine

1998.03

Abstract

The TDMT generation component has as its input a list of words grouped and annotated with the help of linguistic markers. These linguistic markers in general indicate the grammatical function of the respective words, and in theory the task of generation is reduced to inflection and grammatical ordering of the output.
However, the linguistic markers are assigned somewhat mathematically by rules selected on the basis of semantic closeness calculations, often resulting in erroneous structure predictions. To improve the output quality, it would be desirable to have some form of repair strategies for ill-formed input to the generation component, restructering the information given in a sensible way.
As it turns out, this task is not easily achieved without being more careful about the choice of linguistic markers in the transfer component, since the current inconsistencies, irrelevant as they are in the present system, lead to a lot of over-generation even with simple repair strategies.
In this paper I will present some ideas on methods for repair, and why and how they will or will not work, given appropriate input from the transfer component. I will suggest a few changes to the information encoded in the linguistic markers during transfer, providing more structural information to base the repair strategies on, and thus avoiding most causes for over-generation.

ATR Interpreting Telecommunications Research Laboratories

Contents

1 Introduction

TDMT (Transfer-Driven Machine Translation) is an example-based machine translation system, where pretranslated sentence pairs are used to predict the most likely translation for unseen data. As might be expected, the performance of such a system gets better as the size of the training data goes up.

For the Japanese-German component, there are comparably few training sentences, resulting in a relatively high number of unintelligable translations. Whilst quite a few of these are beyond repair, there are some that seem to contain all information needed for a reasonably good translation, provided the structure is slightly adapted. This can be done by using the structural information available, as long as this information is introduced in a consistent way.

During my time here, I have found quite a few sources for serious problems in the system. I have listed these in some 160 bugreports, annotated as to which part of the system they concern. Most of the things included in these bugreports are not referred to in this document, since they are mainly very specific suggestions for changing single rules, or generation strategies. Unfortunately, I could not do these changes myself, since at this time the German rulebase is undergoing drastic changes. However, they are available for reference from Michael Paul (paul@itl.atr.co.jp), and hopefully some of the phenomena listed will already have been eliminated in the process of changing the rulebase.

In the following report I will give some ideas on what repair strategies should work given the appropriate structural information. I will also show why using verbal subcategorization frames is of not much use during the generation stage.

2 The Interface

The TDMT generation component has as its input a list of words grouped and annotated with the help of linguistic markers. These linguistic markers in general indicate the grammatical function of the respective words, and in theory the task of generation is reduced to inflection and grammatical ordering of the output.

Thus, as an example, the sentence

(1) オンドル部屋がよろしいですか

is translated into the following structure during the transfer stage:

```
((FRAGE (SUB (PERSONALPRONOMEN sie PERSON 2A))
        (VP (AP+ (ADVERB gerne))
            (VERB haben) TENSE KONJUNKTIV-2)
        (AKK-OBJ (NP (NOMEN zimmer)
                     (PP (PRAEPOSITION mit)
                         {OHNE}
                         (EIGENNAME (FIX-CAP ondol) GENDER NTR)
                         CASE DAT)))))
```

Encoded in this structure is the information, that the target expression should consist of a single sentence of type question, as indicated by the linguistic marker FRAGE, with the three components subject (SUB), verb (VP), and accusative object (AKK-OBJ). Both the verb and the accusative object are made up of a complex structure defined as verb phrase (VP) and noun phrase (NP) respectively.[1] The PP is encapsulated in the NP containing the noun it modifies, thus disambiguating its scope.

Given the above structure, the generation task is reduced to reordering and inflecting the words. As a rule, in German yes-no-questions the verb is fronted, yielding the correct translation:

(2) *Hätten Sie gerne ein Zimmer mit Ondol?*

'Would you like a room with an ondol?'

Subordinate clauses are marked by the introductory marker SATZGEFUEGE followed by their type, and placed either within the scope of whatever they modify (eg. relative clauses modifying a noun, cf. example 3) or on the top level of the sentence (cf. example 5).

(3) セントラルパークに面しているホテルはありますか

```
((FRAGE (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
        (SUB (PERSONALPRONOMEN es))
        (VERB gibt)
        (AKK-OBJ (NOMEN hotel)
                 (SATZGEFUEGE REL-S
                              (SUB (LC-EXP (RELATIVPRONOMEN der/die/das)))
                              (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
                              (VERB liegen)
                              (PLACE (PRAEPOSITION an)
                                     (EIGENNAME (FIX-CAP central)
                                                (NOMEN park) GENDER MAS)
                                     CASE DAT)))))
```

[1] Note that the names of the linguistic markers used in TDMT do not necessarily correspond to their usual use in standard grammar. Especially, apart from the standard prepositional phrases, anything marked as a *phrase* serves mainly as a means for bracketing words that are in some form related. There is nothing related to the concept of *phrase structure* in TDMT.

(4) *Gibt es ein Hotel, das am Central Park liegt?*

'Is there a hotel that is near Central Park?'

(5) このフェリーが欠航になった場合お金は戻ってくるのですか

```
((FRAGE (*SUB (PARTIKEL man))
        (SUB (NOMEN geld))
        (ATTR (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
              (VERB zurueckbezahlen) VOICE PASSIV)
        (SATZGEFUEGE KOND-S
                     (INTRO (SUBORD-KONJUNKTION wenn))
                     (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
                     (SUB (NP (LC-EXP (DETERMINATIV dieser/diese/dieses))
                              (NOMEN faehre)))
                     (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
                     (VERB ausfallen))))
```

(6) *Wird das Geld zurückbezahlt, wenn diese Fähre ausfällt?*

'Is the money paid back in case the ferry is cancelled?'

To generate two separate sentences, the marker SATZREIHE is used.

(7) とても気分が悪いので医者を呼んでください

```
((ADVERB sehr)
 (SUB (PERSONALPRONOMEN ich))
 (VP (REFLEXIVPRONOMEN mich CASE AKK)
     (VERB fuehlen))
 (ADJEKTIV unwohl)
 (SATZREIHE (IMP-S-I (INTRO (PARTIKEL bitte))
                     (*SUB (PERSONALPRONOMEN sie PERSON 2A))
                     (VERB rufen)
                     (AKK-OBJ (NOMEN arzt)))))
```

(8) *Ich fühle mich sehr unwohl. Bitte rufen Sie einen Arzt.*

'I'm feeling very ill. Please call a doctor.'

Note that the above examples contain default subjects (*SUB) introduced in various places to cope with the frequently occuring ommission of subjects in Japanese. If there is a proper subject present, it will override the defaults, otherwise the first default subject is chosen.

3 Simple Repair Strategies

As we have shown in the previous section, the generation task is more or less trivial given enough information about the structure of the target sentence. However, the TDMT system is based purely on structure alignment of pre-translated example sentences, where the rules to be used for the translation of any sentence are chosen by semantic distance calculations. However, no grammar or semantic formalism is used to determine the target structure. This means, that the generation component is frequently confronted with incomplete or incorrect structures that cannot be generated in a straightforward manner. Whilst some of these erroneous structures are beyond repair, quite a few of them contain all information needed for an intelligable translation, provided the structure is rearranged slightly before generation.

In the following, we will present some ideas on repair strategies that would enable the generation module to produce grammatical sentences even from erroneous structures.

3.1 Trapped Constituents

One of the main preconditions for any repair strategy to work is regularity. The aim should be to provide ways to deal with system-inherent structural problems, leaving patches for particular sentences to the example base in the transfer component.[2]

One of the most frequent and inadvertible errors is one part of the structure getting trapped inside another part, as explicated in example 9.

(9) その近くの郵便局の前の電話ボックスにいるんですよ

```
((HILFSVERB sein)
 (PLACE (PRAEPOSITION in)
        (NP (PP (PRAEPOSITION vor)
                (NP (NOMEN postamt)
                    (NP {BESTIMMT}
                        (PP (PRAEPOSITION in)
                            (NOMEN naehe)
                            CASE DAT)
                        CASE GEN))
                CASE DAT)
            (NOMEN telefonzelle))
        CASE DAT))
```

[2] Most of the problems associated with a single sentence from open data are due to the rather small example base in the German system. It does not make any sense to introduce repair strategies for very specific phenomena that could just as well be covered by a new rule in the transfer component.

(10) *Ist in vor dem Postamt in der Nähe der Telefonzelle.*

'Is in in front of the post office nearby the phone box.'

Here, the PP *vor dem Postamt in der Nähe* got trapped inside the PP *in der Telefonzelle*, making the utterance incomprehensible. However, assuming that an NP grouping a noun and a PP together indicates that the noun is modified by that PP, we can apply the knowledge that in German noun modifying PPs appear after the noun. Thus switching the two constituents puts them into the correct order:

```
(PLACE (PRAEPOSITION in)
       (NP (NOMEN telefonzelle)
           (PP (PRAEPOSITION vor)
               (NP (NOMEN postamt)
                   (NP {BESTIMMT}
                       (PP (PRAEPOSITION in)
                           (NOMEN naehe)
                           CASE DAT)
                       CASE GEN))
               CASE DAT))
       CASE DAT)
```

(11) *in der Telefonzelle vor dem Postamt in der Nähe.*

'in the phone box in front of the post office nearby'

Similarly, in the following example, the PP modifying the noun *Tour* got trapped inside an NP that combines two nouns into one expression.

(12) 伏見桃山宇治コースのチケットがありますか

```
((FRAGE
  (*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
  (AKK-OBJ (NP (NP (NOMEN tour)
                   (PP (PRAEPOSITION fuer)
                       {OHNE}
                       (NP {OHNE}
                           (EIGENNAME+ (FIX-CAP fushimimomoyama) GENDER NTR)
                           (EIGENNAME+ (FIX-CAP uji) GENDER NTR))
                       CASE AKK))
               (NOMEN ticket)))
  (SUB (PERSONALPRONOMEN es)
  (VERB gibt)))
```

(13) *Gibt es eine Tour für Fushimimomoyama Uji Ticket?*

'Is there a tour for Fushimimomoyama Uji ticket?'

In this case we cannot solve the problem by simply reshuffeling the order of the constituents, for the correct translation requieres the noun *Ticket* to go inbetween the noun *Tour* and its modifying PP *für Fushimimomoyama Uji*, thus breaking into the NP grouping these together.

However, this structural change is not as arbitrary as it might seem. It is caused by the differences in the structure of the Japanese input and its proposed German equivalences. During the alignment process, first the Japanese compound noun 伏見桃山宇治コース is translated into an NP structure containing the noun *Tour* and a modifying PP, where the noun コース (*Tour*) is the head of the structure.

$$\text{伏見桃山宇治 <CN-CN> コース} \quad \Longrightarrow \quad (\text{NP Y (PP } \textit{für}\ \text{X))}$$

Making use of this head information, コースのチケット is then considered in the next step, this time yielding a complex noun, *Tour-Ticket*, in German.

$$\text{コース <の> チケット} \quad \Longrightarrow \quad (\text{NP Y Z})$$

Since the entire structure headed by コース is inserted in its place, the modifying PP interferes with the complex noun, separating the two parts during linear generation.

$$\begin{gathered}(\text{(伏見桃山宇治 <CN-CN> コース) <の> チケット}) \\ \Longrightarrow \\ (\text{NP (NP Y (PP } \textit{für}\ \text{X)) Z})\end{gathered}$$

Since TDMT always views a window of only up to three constituents at a time and puts them together to form a new component with one of them as head, this sort of interference of structures is likely to occur frequently and without any sensible cure within the transfer system. However, these phenomena are regular enough to be handled by a repair mechansim in the generation module.

Any two constituents that stand in a certain relationship to each other are grouped together as a complex structure, which should be marked according to the category of its head. For example, the complex structure consisting of a noun that is modified by a prepositional phrase is marked as an NP. Using the head information of complex sturctures already build up, this combining process is continued until the entire sentence is covered by one structure.

In generation, the knowledge of the type of heads used for building the structure can help recover the original intent when a particular sub-structure was introduced. For instance, take the structure introduced in the above example:

(14)
```
(NP (NP (NOMEN tour)
        (PP ...))
    (NOMEN ticket))
```

The inner NP represents a head noun, *Tour*, modified by a PP. The outer NP, however, is really a representation of combining the head of the inner NP, *Tour*, with the noun *Ticket*, resulting in the compound noun *Tour-Ticket*. Whilst compounds cannot be easily separated, modifiers of any of their parts can be taken as modifiers of the whole. We can thus introduce a repair strategy, that if any NP contains only nouns and one or more NPs, the head nouns should be compounded as usual, and all the modifiers applied to the new compound.

From this we get the following transformation rule:

```
(NP (NP (NOMEN tour)
        (PP (PRAEPOSITION fuer)
            (NP ...)))
    (NOMEN ticket))
```

$\Longrightarrow$

```
(NP (NP (NOMEN tour)
        (NOMEN ticket))
    (PP (PRAEPOSITION fuer)
        (NP ...)))
```

Note that the proposed transformation only switches the scope of the two NP markers without changing their internal structure.[3]

In the current example, this would give us the improved translation:

(15) *Gibt es ein Tour Ticket für Fushimimomoyama Uji?*

'Is there a tour ticket for Fushimimomoyama Uji?'

Note that this translation does not have exactly the same meaning as the original Japanese sentence, but it is close enough that it can be understood correctly. A more faithful translation would keep the Japanese structure and read:

(16) *Gibt es ein Ticket für die Fushimimomoyama Uji Tour?*

'Is there a ticket for the Fushimimomoyama Uji Tour?'

Even though this might seem to be yet another word order variation of example 13, it is in fact based on a completely different syntactic structure, with *Fushimimomoyama Uji* and *Tour* forming a compound inside the PP modifying *Ticket*.

Thus, the following transformation is illegal, since it changes the internal structure of the constituents rather than just the scope.

[3] In formal grammar, what is marked here as NOMEN and NP would actually be of the same category.

```
(NP (NP (NOMEN tour)
        (PP (PRAEPOSITION fuer)
            (NP ...)))
    (NOMEN ticket))
```

$\not\Longrightarrow$

```
(NP (NOMEN ticket)
    (PP (PRAEPOSITION fuer)
        (NP (NP ...)
            (NOMEN tour))))
```

Incidentally, it is just a matter of coincidence that the translation equivalents of the two structures came out in reverse, and there is absolutely no way generation can offer repairs for the wrong choice of rule in the transfer component.

3.2 Misassignments

One of the most frequently misplaced or misassigned sentence elements is the linguistic marker SUB introduced by the Japanese topicalization marker は. This has the disadvantage, that the contents of a SUB element should always be checked before considering it as a potential subject. This imposes extra processing on all subjects, even though most of them are correctly assigned.

In the current system, this checking is not carried out, causing default subjects to be thrown away in favour of SUB elements that are not even nouns and thus cannot possibly function as a subject. This is shown in the following examples, where translation variant (a) states the current output, whilst (b) indicates the output that could be achieved by applying consistency checks.

(17) こちらは五万円でございます

```
((*SUB (LC-EXP (PERSONALPRONOMEN er/sie/es)))
 (VERB kostet)
 (SUB (PLACE hier))
 (*HILFSVERB sein)
 (NP (KARDINALZAHL 50000)
     (UNIT yen)))
```

(18) (a) *Hier kostet fünfziglausend Yen.*
'Here costs fiftythousand yen.'

(b) *Hier kostet es fünfziglausend Yen.*
'Here it costs fiftythousand yen.'

(19) 所要時間は六時間になります

```
((SUB (PP (PRAEPOSITION laut)
          {OHNE}
          (NOMEN fahrplan)
          CASE DAT))
 (*SUB (LC-EXP (PERSONALPRONOMEN es)))
 (*VERB dauern)
 (NP (KARDINALZAHL 6)
     (NOMEN stunde)))
```

(20) (a) *Laut Fahrplan dauert sechs Stunden.*
'According to the timetable takes six hours.'

(b) *Laut Fahrplan dauert es sechs Stunden.*
'According to the timetable, it takes six hours.'

Note that the actual source of the misassignment of the SUB marker has been accounted for in the translation by fronting the element, thus moving it into the topic position in German.

4 Consistency

In the previous section, we listed a number of repair strategies to cope with unavoidable errors in the structure. Rather than being exhaustive, this list is meant to exemplify a way the structural information introduced in transfer can be used in a constructive way within the generation component. The structure that is given as input from the transfer process is all the information that is available for generation. If any part of the structure is ungrammatical, generation has no other option than making an educated guess, using all information available to find a plausible explanation for the error and repair it.

However, a vital precondition for such a process is, that the structure assigned during transfer should be consistent. Above all, great care should be taken with the notation used, making sure that there are enough, but not too many distinctions according to the linguistic properties of the constituents. It should be needless to say, that it only makes sense to introduce markers that distinguish constituent types, if they are to be used consistently and carry some meaning.

4.1 Constituent Markers

Both of these conditions do not apply in the current TDMT system. Particularly the marker NP is highly overloaded, basically functioning as a default marker for combining any two constituents to form one. In quite a few cases, there is no intuitive meaning whatsoever associated with the use of

the marker NP, making it difficult to use it for repair strategies in the above mentioned form. For example, the NP marker is also used in some rule to group together two PPs:

(21) リージェンシィホテルはセントラルパークの近くにあるホテルのことですよね

```
((SUB (EIGENNAME (FIX-CAP regency)
                 (NOMEN hotel)
                  GENDER NTR))
 (*HILFSVERB sein)
 (AKK-OBJ (NOMEN hotel))
 (NP (PP (PRAEPOSITION in)
         {BESTIMMT}
         (NOMEN naehe)
         CASE DAT)
     (PP (PRAEPOSITION von)
         (EIGENNAME (FIX-CAP central)
                    (NOMEN park)
                    GENDER MAS)
         CASE DAT))
 (FIX-END , oder ?))
```

(22) *Das Regency Hotel ist ein Hotel in der Nähe vom Central Park, oder?*
'The Regency Hotel is a hotel near Central Park, isn't it?'

Note that the structure in this example also lacks the information, that the PP construction modifies the AKK-OBJ *Hotel*. In the current system, this doesn't make any difference, since the generated sentence is correct, but this should never be an argument for leaving out part of the structure. If any of these rules are used in a different context, there will be some information missing that might have been the clue to repair. On top of that, it gets very difficult to distinguish incomplete structures that have been left that way on purpose and those where something simply went wrong. In the worst case, a simple repair strategy that should cover a wide range of erroneous structures will end up covering a whole load of those intentionally incomplete structures as well, making the overall system performance go down rather than up.

There are many more of this type of inconsistencies, a subset of which (i.e. those I found by coincidence) are referred to in my bugreports. In my opinion, the performance of TDMT could be improved to quite an extend by taking greater care that for each training sentence, all structural information is encoded in such a way, that it gives the right predictions for other sentences this rule will apply to in open data. To achieve this, it would be sensible to analyze first, what type of errors are inherent to the TDMT framework and what sort of information would be useful for repair strategies. The linguistic markers and their function in the rules should then be defined accordingly.

As a second step, the training data would need to be checked as to whether the structures build up for them conform with these definitions, adjusting the rules where neccessary. This is a lot of work, but it should pay off at the end.

4.2 Lexical Category Markers

In the same way the structural markers are not used to their full capacity, the expressive power of the linguistic markers stating word categories is to some extent wasted. This is most obvious with the marker ADVERB, which seems to be the default for any word that happens to require no further inflections during generation. Whilst this might be a valid choice from the morphological perspective, it has quite a drastic impact on the word ordering choices for different syntactic word categories. Graduation particles like *sehr*, for example, modify adverbs and adjectives rather than verbs, and thus obey different word order regularities than adverbs. If they are given the same syntactic category, there is no easy way to make this distinction.

In addition, there seems to be some confusion in the rules about the difference between the categories VERB-ADD, ADVERB and sometimes even PRAEPOSITION, making it quite difficult to define sensible word order rules on these.

5 On the Use of Verbal Subcategorization Frames

One of the major problems in German generation is case assignment. In English, it usually does not matter all that much in what relation a noun phrase stands to the verb, as long as it is positioned correctly. Since TDMT imposes some word order constraints during the transfer stage, the result still will come out correct most of the times, even if the function of the noun is not known. In German, the positioning of such a noun phrase in a sentence might also be correct, but there won't be any case assigned to it. Since using the wrong case will inevitably lead to great problems in understanding, it is usually better to leave out the article and generate the noun uninflected. The slightly awkward impression any German speaker will get on hearing this will likely induce an inference process to recover the case that noun should have been assigned. Given that the structure is otherwise reasonably correct, this will eventually lead to an understanding of the intended sentence meaning.

However, it would be much nicer to have some mechansim that recovers the case requirements of an unassigned NP and gives the sentence more naturalness. The first idea that comes to mind is to use the subcategorization information associated with the respective verb. If the number of arguments

to the verb and their required cases were known, we could check for their completeness and match any unassigned NPs to slots in the case frame that are not filled. Even better still, semantic subcategorization frames, i.e. a list of the number and semantic type of arguments associated with the verb, could be used to verify the arguments assigned and help choose between multiply assigned arguments or even different translation variants.

5.1 Practical Concerns

The first problem is the acquisition of the verbal case frames. Since the TDMT system doesn't assign the object positions in a reliable way, it cannot be used for extracting case frames even for the training data. Extracting them by hand is a tedious task that is bound to lead to inconsistent results. On top of that, especially when integrating semantic information, the database will be incomplete, making it hard to decide whether a suitable case frame cannot be found because the current assignments are simply ungrammatical, or because the respective entry happens to be missing from the database. We would have to use the same training data as for the rules, thus only covering those cases that are likely to come out correct in most cases anyway.

For German, there exists a database, CELEX, that contains nearly all syntactic case frames for a large number of German verbs, stating the presence or absence of different complement types (i.e. one or more accusative objects, dative objects, prepositional phrases, etc.). However, this database is not in all parts correct and covers case frames for many different readings of a particular word. Thus in most cases it will be a matter of lucky guessing whether we pick the correct case frame out of the many.

As said above, which case frame is applicable, be it syntactic or semantic, depends to a large part on the intended reading of the verb. For two languages as different as Japanese and German, quite frequently the different readings of some word in one language have entirely different translations in the other. Thus, using knowledge about the source language expression might help to pick the applicable case frame in the target language. However, this is only possible, if some form of marking takes place during transfer, since there is no way for generation to retrieve this information otherwise.

5.2 Syntactic Case Frames

As a case study, we tested the usefulness of syntactic case frame information for assigning case to noun phrases in the structure that did not have any particular function assigned to them. For our case study, we used the CELEX database, and compared the entries from the database with the number and type of arguments found in the structure. If any of the CELEX frames included a nominative complement, the noun phrase would be moved to the

nominative complement position. Otherwise, if all CELEX entries for that verb required an accusative object, it would be moved to that position.

The assignment of nominative complements improved the translation quality in quite a few cases, making the sentences more natural by inserting articles, as shown in the following example:

(23) 一番近いところで釜山港ですね

(a) *Der nächste Ort ist Hafen Pusan.*
'The next place is harbour Pusan.'

(b) *Der nächste Ort ist* der *Hafen Pusan.*
'The next place is the harbour Pusan.'

However, this could also have been achieved in most cases by adding the appropriate subcategorization information through the respective rules during the transfer stage.

There were also cases, where the translation quality deteriorated because of these reassignements, most notably in cases where the structure was inconsistent, as for instance in example 21, repeated here as 24:

(24) リージェンシィホテルはセントラルパークの近くにあるホテルのことですよね

(a) *Das Regency Hotel ist ein Hotel in der Nähe vom Central Park, oder?*
'The Regency Hotel is a hotel near Central Park, isn't it?'

(b) *Das Regency Hotel ist in der Nähe vom Central Park ein Hotel, oder?*
'The Regency Hotel is near Central Park a hotel, isn't it?'

Recall that the two PPs *in der Nähe* and *vom Central Park* were grouped together as an NP, which has in this case been taken to be the recovered nominative complement, and thus moved in front of the accusative object *ein Hotel.*

The recovery of accusative objects was not very successful with respect to the intended effect. However, it did turn out to be quite a good debugging tool, uncovering quite a lot of problematic rules from transfer. Most of these phenomena can be found in my bugreports.

What yielded very good results was the simple replacement of a default accussative object in the structure by any unassigned NP. This got rid of most spurious personal pronouns in object position, as shown in the following example:

(25) 何か不都合なことでもございましたか

```
((FRAGE (SUB (PERSONALPRONOMEN sie PERSON 2A))
        (AP+ (ADVERB noch))
        (*SUB (PERSONALPRONOMEN ich))
        (*SUB wir)
        (HILFSVERB haben)
        (*AKK-OBJ (PERSONALPRONOMEN es))
        (NP (DETERMINATIV-INDEF irgendein)
            (NOMEN problem))))
```

(26) (a) *Haben Sie es noch irgendein Problem?*
'Do you have it any other problem?'

(b) *Haben Sie noch irgendein Problem?*
'Do you have any other problem?'

This strategy could possibly also be used for encoding subcategorization information at the transfer stage.

5.3 Problems

From this it can be seen, that using subcategorization information at the generation stage is of not much use in the current system. There are too many inconsistencies interfereing, making the overall performance go down rather than up.

Especially with semantic case frames, there would also remain the question about what could actually be done if some parts of the structure were of the wrong type. As said above, the database would most likely be incomplete, making it near to impossible to decide whether the structure is wrong or just unknown. In addition, in cases where some case assignment has gone completely wrong in the rules, the chances of actually finding an understandable translation that preserves the intended meaning are rather low. It is questionable, whether such a result is more desirable than something that cannot be understood at all. For this reason, it might in fact be safer to leave the structure as it is.

There is not much scope for the subcategorization information actually helping to choose the best translation out of multiple candidates with equal score, since they rarly differ with respect to the argument assignment.

6 Conclusions and Outlook

As should have become clear from this report, the quantity and quality of the structural information assigned to constituents during transfer is of vital

importance for the translation quality. If generation was going to make better use of this information, a lot of errors could be eliminated by very simple structural repairs.

To improve the reliability of the linguistic markers assigned, we have to make sure, that the rules associated with each training sentence produce complete and correct structures. It is important that all structural dependencies should be marked as such.

But most important of all, no rules should be taylored just for the sake of making one sentence work. When writing rules, this should always be done in a way that is as general as possible, with as little harm as possible done by the rule if applied in the wrong context.

Maybe, it would be worth introducing some form of "generality weights", indicating how specific to the particular semantic context a rule is. This way, we could also give more importance to one side of the pattern to match by assigning different weights to the semantic distance calculations of both sides.

Even though the use of subcategorization information turned out to be of little use at the generation stage, it might be useful to test its potentials at the transfer stage. Frequently, a verb has different case frames depending on its reading. This information cannot be recovered at the generation stage.

As said before, for two languages as different as Japanese and German, the different readings of a word in one language often correspond to different translations in the other language. Thus, if we could identify the subcategorization frame of a verb in the source language, we might actually be able to use this information to predict the correct translation in the target language. This could be done by associating each translation variant with a subcategorization frame.

Also, from the knowledge about what semantic type an argument is expected to have, we could predict the translation variant semantically closest to the expected type.